

## PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2001-256253

(43)Date of publication of application : 21.09.2001

(51)Int.Cl.

G06F 17/30

(21)Application number : 2000-069477

(71)Applicant : KDDI CORP

(22)Date of filing : 13.03.2000

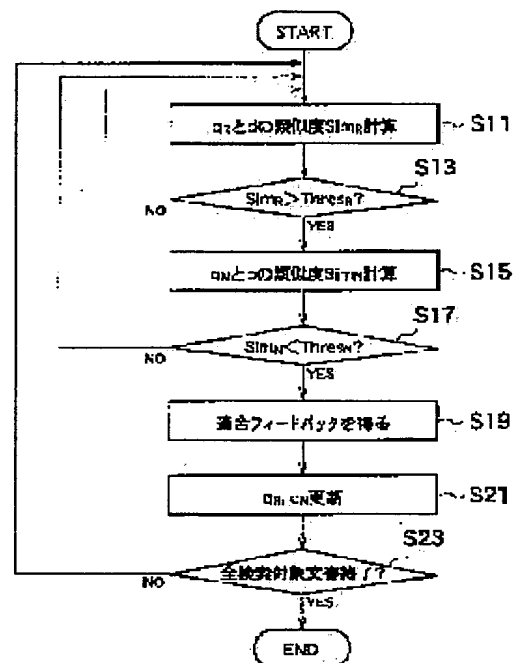
(72)Inventor : HOASHI KEIICHIRO  
INOUE NAOKI  
MATSUMOTO KAZUNORI  
HASHIMOTO KAZUO

## (54) METHOD AND DEVICE FOR FILTERING DOCUMENT

## (57)Abstract:

PROBLEM TO BE SOLVED: To provide a method and a device for filtering document, with which the selection of an unsuited document can be avoided by generating an unsuited profile expressing the features of the erroneously selected unsuited document, and not selecting the document of high similarity to the unsuited profile.

SOLUTION: Similarity SimR between a suited profile qR and a retrieval object document (d) is calculated (step S11), this similarity SimR is compared with a threshold ThresR and when SimR is greater than ThresR, similarity SimN between an unsuited profile qN and the retrieval object document (d) is calculated (step S15). When SimN is greater than ThresN, the retrieval object document is regarded as an unsuited document and excluded. When SimN is smaller, this document is regarded as a suited document and selected, suited feedback is provided and qR and qN are updated (step S21).



## LEGAL STATUS

[Date of request for examination]

10.02.2003

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

Copyright (C); 1998;2003 Japan Patent Office

(19)日本国特許庁 (J P)

(12) 公 開 特 許 公 報 (A)

(11)特許出願公開番号  
特開2001-256253  
(P2001-256253A)

(43)公開日 平成13年9月21日(2001.9.21)

(51)Int.Cl.<sup>7</sup>

G 0 6 F 17/30

識別記号

3 4 0  
1 7 0

F I

G 0 6 F 17/30

テーマコード(参考)

3 4 0 A 5 B 0 7 5  
1 7 0 A

審査請求 未請求 請求項の数4 O L (全 14 頁)

(21)出願番号 特願2000-69477(P2000-69477)

(22)出願日 平成12年3月13日(2000.3.13)

(71)出願人 000208891

ケイディーディーアイ株式会社  
東京都新宿区西新宿二丁目3番2号

(72)発明者 帆足 啓一郎

埼玉県上福岡市大原2-1-15 株式会社  
ケイディディ研究所内

(72)発明者 井ノ上 直己

埼玉県上福岡市大原2-1-15 株式会社  
ケイディディ研究所内

(74)代理人 100083806

弁理士 三好 秀和 (外3名)

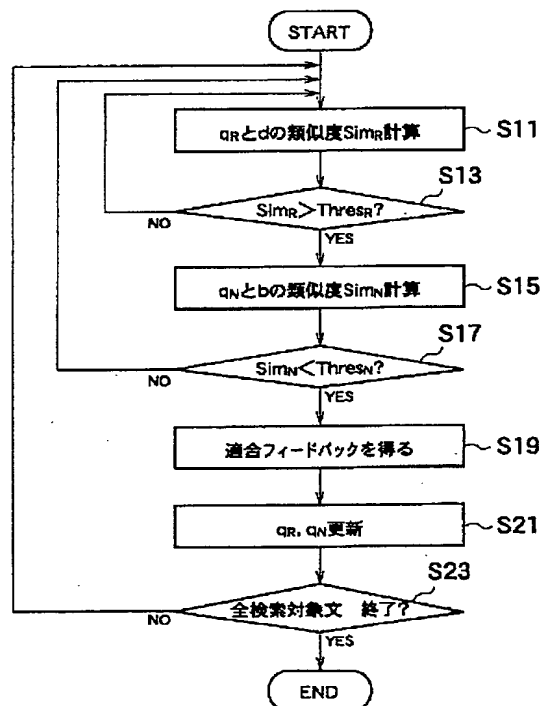
最終頁に続く

(54)【発明の名称】 文書フィルタリング方法および装置

(57)【要約】

【課題】 誤って選択された非適合文書の特徴を表す非適合プロファイルを作成し、非適合プロファイルとの類似度が高い文書を選択しないということで非適合文書の選択を回避し得る文書フィルタリング方法および装置を提供する。

【解決手段】 適合プロファイル $q_R$ と検索対象文書 $d$ との類似度 $Sim_R$ を計算し(ステップS11)、該類似度 $Sim_R$ と閾値 $Thres_R$ と比較し、 $Sim_R$ が $Thres_R$ より大きい場合、非適合プロファイル $q_N$ と検索対象文書 $d$ との類似度 $Sim_N$ を計算し(ステップS15)、 $Sim_N$ が $Thres_N$ より大きい場合、検索対象文書を非適合文書と見なして除外し、小さい場合、適合文書と見なして選択し、適合フィードバックを得て、 $q_R$ 、 $q_N$ を更新する(ステップS21)。



**【特許請求の範囲】**

【請求項 1】 ユーザの要求を表す適合プロファイルに適合する文書を検索対象文書の中から抽出して出力する文書フィルタリング方法であって、前記適合プロファイルに対して類似度が高いが、ユーザの要求に適合しないと判定された非適合文書の特徴づけるプロファイルを非適合プロファイルとして作成し、前記適合プロファイルと検索対象文書との類似度を算出し、この算出した類似度を所定の適合用閾値と比較し、該類似度が所定の適合用閾値より大きい場合、前記検索対象文書と前記非適合プロファイルとの類似度を算出し、この非適合プロファイルとの類似度を所定の非適合用閾値と比較し、該類似度が所定の非適合用閾値より大きい場合、該検索対象文書を非適合文書と見なして除外し、小さい場合、適合文書と見なして選択し、前記非適合文書および適合文書の適合フィードバックを行い、適合プロファイルおよび非適合プロファイルを更新することを特徴とする文書フィルタリング方法。

【請求項 2】 前記非適合プロファイルの更新は、選択文書に出現する単語から単語寄与度に基づき選択文書の特徴を表す単語を抽出し、この抽出された単語の寄与度を算出し、この単語寄与度に重みをかけて、単語に対するスコアを算出し、前記選択文書が適合文書である場合には、前記スコアを非適合プロファイルから減算し、前記選択文書が非適合文書である場合には、前記スコアを非適合プロファイルに加算して更新することを特徴とする請求項 1 記載の文書フィルタリング方法。

【請求項 3】 ユーザの要求を表す適合プロファイルに適合する文書を検索対象文書の中から抽出して出力する文書フィルタリング装置であって、前記適合プロファイルに対して類似度が高いが、ユーザの要求に適合しないと判定された非適合文書の特徴づけるプロファイルを非適合プロファイルとして作成する非適合プロファイル作成手段と、前記適合プロファイルと検索対象文書との類似度を算出する適合プロファイル類似度算出手段と、この算出した類似度を所定の適合用閾値と比較し、該類似度が所定の適合用閾値より大きい場合、前記検索対象文書と前記非適合プロファイルとの類似度を算出する非適合プロファイル類似度算出手段と、この非適合プロファイルとの類似度を所定の非適合用閾値と比較し、該類似度が所定の非適合用閾値より大きい場合、該検索対象文書を非適合文書と見なして除外し、小さい場合、適合文書と見なして選択する選択手段と、前記非適合文書および適合文書の適合フィードバックを

行い、適合プロファイルおよび非適合プロファイルを更新する更新手段とを有することを特徴とする文書フィルタリング装置。

【請求項 4】 前記更新手段は、選択文書に出現する単語から単語寄与度に基づき選択文書の特徴を表す単語を抽出する単語抽出手段と、この抽出された単語の寄与度を算出する単語寄与度算出手段と、この単語寄与度に重みをかけて、単語に対するスコアを算出するスコア算出手段と、前記選択文書が適合文書である場合には、前記スコアを非適合プロファイルから減算し、前記選択文書が非適合文書である場合には、前記スコアを非適合プロファイルに加算して更新するスコア加減算手段とを有することを特徴とする請求項 3 記載の文書フィルタリング装置。

**【発明の詳細な説明】****【0001】**

【発明の属する技術分野】 本発明は、ユーザの要求を表す適合プロファイルに適合する文書を検索対象文書の中から抽出して出力する文書フィルタリング方法および装置に関し、更に詳しくは、適合プロファイルとの類似度は高いが、ユーザの要求に適合しないと判定された非適合文書の特徴づけるプロファイルを非適合プロファイルとして作成し、この非適合プロファイルに対して類似度が高い文書を選択しないように削除して文書のフィルタリングを行う文書フィルタリング方法および装置に関する。

**【0002】**

【従来の技術】 この種の文書フィルタリング技術は、例えばメールによるニュース情報配信サービスなどのように大量のテキスト情報の流れの中からユーザの嗜好に合致した情報のみを抽出して、ユーザに提供するのにも有効である。すなわち、文書フィルタリングとは、次々に流れてくる検索対象文書の中からユーザの要求を満たしている文書のみを取得し、ユーザに提供するタスクである。

【0003】 このような文書フィルタリングにおいて、ユーザの要求はプロファイルとしてフィルタリングシステム内で表されている。そして、フィルタリングシステムは、順次流れてくる検索対象文書 1 つ 1 つについて、このプロファイルを満たしているか否かを判断し、要求を満たしている文書のみをユーザに提示する。ユーザは、提示された文書に対して実際に要求を満たしているか否かを判断し、その判断をフィルタリングシステムにフィードバックする。多くの場合、フィルタリングシステムはユーザからのフィードバックを基にプロファイルを更新することによりフィルタリングの精度向上を図っている。

【0004】 フィルタリングシステムには多くの場合情報検索で使用する技術が適用されている。システムに

入力される各文書やプロファイルはベクトル空間モデルなどに基づいてシステム内で表現され、各文書がプロファイルを満たしているか否かの判断基準として、プロファイルと文書との間の類似度が使用されることが多い。また、文書フィルタリングのプロファイル更新には情報検索における検索式拡張手法を適用する手法が多く用いられている。すなわち、ユーザからの適合フィードバック情報に基づき、選択された文書から抽出された情報をプロファイルに追加して、プロファイルを更新することにより、プロファイルを精緻化させる処理を行うのである。

【0005】このようなプロファイル更新を利用した従来の文書フィルタリング方法の処理手順について図5に示すフローチャートを参照して説明する。

【0006】この文書フィルタリング方法では、まずユーザの要求であるプロファイル $q$ に類似する文書を検索対象文書 $d$ の中から検索すべくプロファイル $q$ と検索対象文書 $d$ の類似度を計算する(ステップS71)。この計算したプロファイル $q$ と検索対象文書 $d$ との類似度が所定の閾値より大きいかなかを判定する(ステップS73)。該類似度が所定の閾値よりも大きくない場合には、最初のステップS71に戻り、次の検索対象文書について同じ処理を繰り返すが、類似度が所定の閾値よりも大きい場合には、適合フィードバックを得て(ステップS75)、プロファイル $q$ を更新する(ステップS7

$$\text{Cont}(w_i, q, d)$$

$$= \text{Sim}(q, d) - \text{Sim}(q'(w_i), d'(w_i))$$

… (1)

ただし、 $\text{Sim}(q, d)$ は、 $q, d$ 間の類似度を表し、 $q'(w_i)$ は $q$ から単語 $w_i$ を除いた入力文、 $d'(w_i)$ は $d$ から単語 $w_i$ を除いた文書を表すとする。

【0011】すなわち、単語寄与度 $\text{Cont}(w_i, q, d)$ とは、 $q$ と $d$ との類似度と単語 $w_i$ が存在しない場合の $q$ と $d$ との類似度との差である。従って、 $q$ と $d$ に出現する全ての単語のうち、類似度を向上させる単語の寄与度は正であり、逆に類似度を下げる単語の寄与度は負である。

【0012】また、文献「帆足、松本、井ノ上、橋本：文書間の類似度における単語寄与度を利用した検索式拡張手法、情報処理学会論文誌：データベース、Vol.40, No. SIG8(TOD4), pp. 63-73, 1999」によれば、出現単語の多くの寄与度は0に近く、類似度に有意な影響を与えている単語は少ない。そのうち、寄与度が大きく正の値を持つ単語は、入力文と検索対象文書の両方に存在した単語

$$\text{Score}(w) = \text{wgt} \times \sum_{d \in D_{\text{rel}}(q)}$$

次に、抽出された単語のうち元の検索式に含まれていない単語を検索式に加えることで検索式拡張を実現する。

【0015】ある単語 $w$ を入力文のベクトルに加える際

7) という処理を全ての検索対象文書について行って、処理を終了する(ステップS79)。

【0007】このような文書フィルタリング方法では、一般的にプロファイルとの類似度が閾値を上回る文書を選択してユーザに提示するという流れで文書のフィルタリングを行っているが、このようなフィルタリング手法では後述するように適切な閾値を設定することが難しく、多くの適合文書を選択するために閾値を低く設定すると、誤って選択される非適合文書の数が大きく増大し、また逆に非適合文書の誤り選択を減少させるために閾値を高く設定すると、多くの適合文書を見逃してしまうことになる。

【0008】このような文書フィルタリング方法におけるプロファイルの更新手法には上述したように情報検索で使用される検索式拡張技術が適用されることが多いが、次に情報検索において高い精度が得られる単語寄与度に基づく検索式拡張手法を適用したプロファイル更新手法について説明する。

【0009】まず、単語寄与度に基づく検索式拡張手法について説明する。単語寄与度とは、文書間の類似度における各単語の影響を数値化した尺度である。ある入力文 $q$ と検索対象文書 $d$ との間の類似度における単語 $w_i$ の単語寄与度を式(1)によって定義する。

【0010】

【数1】

である。一方、大きな負の値の寄与度を持つ単語は一方の文書にのみ存在し、かつ、その文書の特徴を顕著に表す単語であると考えられる。そこで、単語寄与度に基づいた検索式拡張手法では以下のように検索式の拡張を行っている。

【0013】まず、入力文 $q$ と適合している文書群

$$D_{\text{rel}}(q) = \{d_1, \dots, d_{\text{Num}}\} \quad \dots (2)$$

中の各文書に出現する全ての単語の寄与度を求め、各類似文書から単語寄与度の低い単語を $N$ 個抽出する。次に抽出された各単語の寄与度の総和に重み $\text{wgt}$ をかけ、これを単語 $w$ に対するスコアとする。単語 $w$ の入力文 $q$ と文書 $d$ の類似度に対する寄与度を $\text{Cont}(w_i, q, d)$ とすると、単語 $w$ のスコア $\text{Score}(w)$ は式(3)によって表される。

【0014】

【数2】

$$\text{Cont}(w, q, d) \quad \dots (3)$$

には、式(3)で計算されたスコア $\text{Score}(w)$ を単語 $w$ が入力文に出現する頻度(単語出現頻度  $\text{tf}$ )と見なし、入力文のベクトル内で単語 $w$ を表す要素の値

を計算する。ベクトルの各要素が $TF * IDF$ によって計算されている場合、 $Score(w)$ を $tf$ とすることで $TF$ を算出し、更に単語 $w$ の $IDF$ をかけ、その結果得られた $TF * IDF$ 値を入力文のベクトルの単語 $w$ の要素に入れることにより、検索式拡張を行う。

【0016】次に、単語寄与度に基づくプロファイル更新手法について説明する。

【0017】単語寄与度による検索式拡張では、初期検索の結果に対するフィードバックにより得られた適合文書集合中の各文書から寄与度に基づいて抽出された単語の寄与度の総和に重みを掛けることで、各単語に対する

$$Score_{rel}(w_i) = wgt_{rel} \times Cont(w_i, q, d) \quad \dots (4)$$

$$Score_{nrel}(w_i) = wgt_{nrel} \times Cont(w_i, q, d) \quad \dots (5)$$

上記の式によって求めた各単語のスコアを単語出現頻度 $tf$ として扱い、 $TF * IDF$ 法により各単語の重みを算出する。そして、抽出された単語が適合文書中の単語の場合はその単語と重みをプロファイルに加え、非適合文書中の単語の場合は単語と重みをもとのプロファイルから引く。すなわち、適合文書から選択された各単語の要素が元のプロファイルに加えられ、非適合文書から選択された各単語の要素が元のプロファイルから引かれるということである。なお、この処理により負の重みを持った単語は、類似度計算に使用されない。

【0020】以上の処理により、プロファイルを表すベクトルの、値を持たなかった次元が正の値を持つようになり、プロファイルの情報が拡張されることになる。また、適合文書と非適合文書に共起する単語の重みが抑制され、適合文書のみに出現する単語の重みが強調され

$TF \ factor$

$$\log(1 + tf_{ij}) \quad \dots (6)$$

$IDF \ factor$

$$\log\left(\frac{M}{df_j}\right) \quad \dots (7)$$

ただし、 $tf_{ij}$ は文書 $d_i$ 内の単語 $w_j$ の出現頻度、 $df_j$ は単語 $w_j$ が出現する文書数、 $M$ は語彙作成時に使用された文書集合に含まれる文書数とする。

【0024】また、類似度は式(8)で定義されるプロ

$$\cos(\vec{q}, \vec{d}) = \frac{\vec{q} \cdot \vec{d}}{|\vec{q}| |\vec{d}|} \quad \dots (8)$$

ただし、 $\vec{q}$ 、 $\vec{d}$ はそれぞれ入力文書と検索対象文書を表すベクトルとし、 $|\vec{d}|$ は $\vec{d}$ のユークリッド長とする。

スコアを求めた。ここでは、フィルタリング中に選択された文書1つ毎に単語寄与度に基づき文書中から単語を抽出し、その情報を直前のプロファイルに加えることで、随時プロファイルを更新する。

【0018】まず、選択された文書が適合文書の場合には、抽出した単語 $w_i$ に対するスコア $Score_{rel}(w_i)$ を式(4)により算出し、選択した文書が非適合文書中の場合には、抽出した単語 $w_i$ のスコア $Score_{nrel}(w_i)$ を式(5)により算出する。

【0019】

【数3】

る。

【0021】ここでは検索対象文書とプロファイルの両方をベクトル空間モデルを用いて表し、両者間の類似度を計算することで各文書に対するフィルタリングを実現した。

【0022】ベクトル空間モデルで表現する際、各文書およびプロファイルを表すベクトルの各要素の重みは $TF * IDF$ 法により算出する。ここでは、最も有効な情報検索システムの1つであるSMARTにおいて使用されているアルゴリズムに基づき $TF$ および $IDF$ の計算式を使用した。本実験で使用した $TF$ および $IDF$ の計算式を式(6)、式(7)に示す。

【0023】

【数4】

ファイルと検索対象文書のベクトルのコサイン値を取ることで正規化した値として求めた。

【0025】

【数5】

本プロファイル更新手法では、式(4)、(5)によって求めた各単語のスコアを単語出現頻度  $t f$  として扱い、式(6)、(7)によりTFおよびIDFを計算する。従って、各単語のTF\*IDF値は式(9)、(1

0)により算出される。

【0026】

【数6】

$$\text{Value}_{\text{rel}}(w_i) = \log(1 + \text{Score}_{\text{rel}}(w_i)) \times \log\left(\frac{M}{d f_i}\right) \quad \dots (9)$$

$$\text{Value}_{\text{nrel}}(w_i) = \log(1 + \text{Score}_{\text{nrel}}(w_i)) \times \log\left(\frac{M}{d f_i}\right) \quad \dots (10)$$

ただし、 $d f_i$ は単語 $w_i$ が出現する文書数、 $M$ は $d f$ のリストの作成に使用された文書数とする。

【0027】また、プロファイル $q$ および文書 $d$ をベクトル空間モデルによってそれぞれ式(11)、(12)のように表すとする。

【0028】

【数7】

$$\vec{q} = (q_1, \dots, q_n) \quad \dots (11)$$

$$\vec{d} = (d_1, \dots, d_n) \quad \dots (12)$$

ただし、 $q_1, \dots, q_n$ はプロファイル中の単語の重みを、 $d_1, \dots, d_n$ は各文書中の単語の重みを表し、 $n$ はベクトルの次元数を表す。

【0029】更新後のプロファイルを式(13)で表すとする、抽出された各単語 $w_i$ について、適合文書中の単語の場合は式(14)により表され、非適合文書中の単語の場合は式(15)により表される。

【0030】

【数8】

$$\vec{q}_{\text{new}} = (q_1', \dots, q_n') \quad \dots (13)$$

$$q_i' = q_i + \text{Value}_{\text{rel}}(w_i) \quad \dots (14)$$

$$q_i' = q_i - \text{Value}_{\text{nrel}}(w_i) \quad \dots (15)$$

すなわち、適合文書から選択された各単語の要素が元のプロファイルの要素に加えられ、非適合文書から選択された各単語の要素が元のプロファイルの要素から引かれるということである。なお、この処理により負の重みを持った単語は、類似度計算に使用されない。

【0031】図6は、上述したプロファイル更新手法の手順を示すフローチャートである。同図に示すように、ベクトル空間モデルによって表されたプロファイル $q$ および更新後のプロファイル $q_{\text{new}}$ 、および選択文書 $d$ に対して、まず選択文書 $d$ から単語集合 $W$ を抽出し(ステップS83)、選択文書 $d$ が適合文書であるか否かを判定する(ステップS85)。選択文書 $d$ が適合文書である場合には、各単語に対するスコアを式(4)のように算出し(ステップS87)、また(14)で示したよう

に各単語 $w_i$ のスコアをプロファイル $q$ に加える(ステップS89)。また、選択文書 $d$ が適合文書でない場合には、各単語に対するスコアを式(5)のように算出し(ステップS88)、また(15)で示したように各単語 $w_i$ のスコアをプロファイル $q$ から引き(ステップS91)、このような加減算したプロファイル $q$ を更新後のプロファイル $q_{\text{new}}$ として設定する(ステップS93)。

【0032】上述したプロファイル更新手法に基づき、TREC-8のFiltering Trackで用意されたデータを使用した評価実験を行った。図7に類似度の閾値を0.1に設定した場合に選択された適合文書および非適合文書のプロファイルとの類似度を示す。

【0033】図7から閾値を大きく超える類似度を有す

る非適合文書は少ないものの、閾値近辺の類似度では適合文書と非適合文書が混在していることがわかる。

#### 【0034】

【発明が解決しようとする課題】上述したように、従来の文書フィルタリング方法では、閾値を大きく超える類似度を有する非適合文書は少ないものの、閾値近辺の類似度では適合文書と非適合文書が混在していて、これらの類似度を有する適合文書のみを選択することは不可能であり、例えば閾値を低く設定して、多くの適合文書を選択しようすると、誤って選択される非適合文書が増大し、また逆に閾値を高く設定すると、誤って選択される非適合文書は減少するが、適合文書の数も減少してしまうという問題がある。

【0035】すなわち、情報検索における検索式拡張手法をプロフィール更新に適用し、単純に類似度の閾値を設定することにより多くの適合文書を取得しようすると、非適合文書も多く選択されてしまうという問題がある。

【0036】本発明は、上記に鑑みてなされたもので、その目的とするところは、誤って選択された非適合文書の特徴を表す非適合プロフィールを作成し、非適合プロフィールとの類似度が高い文書を選択しないということで非適合文書の選択を回避し得る文書フィルタリング方法および装置を提供することにある。

#### 【0037】

【課題を解決するための手段】上記目的を達成するため、請求項1記載の本発明は、ユーザの要求を表す適合プロフィールに適合する文書を検索対象文書の中から抽出して出力する文書フィルタリング方法であって、前記適合プロフィールに対して類似度が高いが、ユーザの要求に適合しないと判定された非適合文書の特徴づけるプロフィールを非適合プロフィールとして作成し、前記適合プロフィールと検索対象文書との類似度を算出し、この算出した類似度を所定の適合用閾値と比較し、該類似度が所定の適合用閾値より大きい場合、前記検索対象文書と前記非適合プロフィールとの類似度を算出し、この非適合プロフィールとの類似度を所定の非適合用閾値と比較し、該類似度が所定の非適合用閾値より大きい場合、該検索対象文書を非適合文書と見なして除外し、小さい場合、適合文書と見なして選択し、前記非適合文書および適合文書の適合フィードバックを行い、適合プロフィールおよび非適合プロフィールを更新することを要旨とする。

【0038】請求項1記載の本発明にあつては、適合プロフィールに対して類似度が高いが、ユーザの要求に適合しないと判定された非適合文書の特徴づけるプロフィールを非適合プロフィールとして作成し、適合プロフィールと検索対象文書との類似度が所定の適合用閾値より大きい場合、検索対象文書と非適合プロフィールとの類似度を算出し、この類似度が所定の非適合用閾値より大

きい場合、検索対象文書を非適合文書と見なして除外し、小さい場合、適合文書と見なして選択し、適合フィードバックを行い、適合プロフィールおよび非適合プロフィールを更新するため、適合文書とともに選択される非適合文書の数少なくすることができ、フィルタリング性能を向上することができる。

【0039】また、請求項2記載の本発明は、請求項1記載の発明において、前記非適合プロフィールの更新が、選択文書に出現する単語から単語寄与度に基づき選択文書の特徴を表す単語を抽出し、この抽出された単語の寄与度を算出し、この単語寄与度に重みをかけて、単語に対するスコアを算出し、前記選択文書が適合文書である場合には、前記スコアを非適合プロフィールから減算し、前記選択文書が非適合文書である場合には、前記スコアを非適合プロフィールに加算して更新することを要旨とする。

【0040】請求項2記載の本発明にあつては、選択文書の特徴を表す単語を抽出し、この抽出された単語の寄与度を算出し、この単語寄与度に重みをかけて、単語に対するスコアを算出し、選択文書が適合文書である場合には、スコアを非適合プロフィールから減算し、前記選択文書が非適合文書である場合には、前記スコアを非適合プロフィールに加算して更新するため、このように更新される非適合プロフィールを使用することにより文書フィルタリング精度を向上することができる。

【0041】更に、請求項3記載の本発明は、ユーザの要求を表す適合プロフィールに適合する文書を検索対象文書の中から抽出して出力する文書フィルタリング装置であって、前記適合プロフィールに対して類似度が高いが、ユーザの要求に適合しないと判定された非適合文書の特徴づけるプロフィールを非適合プロフィールとして作成する非適合プロフィール作成手段と、前記適合プロフィールと検索対象文書との類似度を算出する適合プロフィール類似度算出手段と、この算出した類似度を所定の適合用閾値と比較し、該類似度が所定の適合用閾値より大きい場合、前記検索対象文書と前記非適合プロフィールとの類似度を算出する非適合プロフィール類似度算出手段と、この非適合プロフィールとの類似度を所定の非適合用閾値と比較し、該類似度が所定の非適合用閾値より大きい場合、該検索対象文書を非適合文書と見なして除外し、小さい場合、適合文書と見なして選択する選択手段と、前記非適合文書および適合文書の適合フィードバックを行い、適合プロフィールおよび非適合プロフィールを更新する更新手段とを有することを要旨とする。

【0042】請求項3記載の本発明にあつては、適合プロフィールに対して類似度が高いが、ユーザの要求に適合しないと判定された非適合文書の特徴づけるプロフィールを非適合プロフィールとして作成し、適合プロフィールと検索対象文書との類似度が所定の適合用閾値より大

大きい場合、検索対象文書と非適合プロフィールとの類似度を算出し、この類似度が所定の非適合用閾値より大きい場合、検索対象文書を非適合文書と見なして除外し、小さい場合、適合文書と見なして選択し、適合フィードバックを行い、適合プロフィールおよび非適合プロフィールを更新するため、適合文書とともに選択される非適合文書の数少なくすることができ、フィルタリング性能を向上することができる。

【0043】請求項4記載の本発明は、請求項3記載の発明において、前記更新手段が、選択文書に出現する単語から単語寄与度に基づき選択文書の特徴を表す単語を抽出する単語抽出手段と、この抽出された単語の寄与度を算出する単語寄与度算出手段と、この単語寄与度に重みをかけて、単語に対するスコアを算出するスコア算出手段と、前記選択文書が適合文書である場合には、前記スコアを非適合プロフィールから減算し、前記選択文書が非適合文書である場合には、前記スコアを非適合プロフィールに加算して更新するスコア加減算手段とを有することを要旨とする。

【0044】請求項4記載の本発明にあつては、選択文書の特徴を表す単語を抽出し、この抽出された単語の寄与度を算出し、この単語寄与度に重みをかけて、単語に対するスコアを算出し、選択文書が適合文書である場合には、スコアを非適合プロフィールから減算し、前記選択文書が非適合文書である場合には、前記スコアを非適合プロフィールに加算して更新するため、このように更新される非適合プロフィールを使用することにより文書フィルタリング精度を向上することができる。

【0045】

【発明の実施の形態】以下、図面を用いて本発明の実施の形態を説明する。図1は、本発明の一実施形態に係る文書フィルタリング方法を実施する文書フィルタリング装置の構成を示すブロック図である。同図に示す文書フィルタリング装置は、ユーザの要求であるプロフィールや検索対象文書などを入力する入力部1と、適合プロフィールとの類似度を算出する適合プロフィール類似度算

出部3aおよび非適合プロフィールとの類似度を算出する非適合プロフィール類似度算出部3bからなる適合性判定部3と、ユーザの要求に適合する文書をユーザに出力する出力部5と、プロフィールなどの情報を記憶する記憶部7と、プロフィールを更新するプロフィール更新部9と、適合フィードバックを行うフィードバック部11とから構成されている。

【0046】次に、上述したように構成される本実施形態の文書フィルタリング装置の作用を説明する前に、本発明の文書フィルタリング方法の概要について説明する。上述したように、従来使用されてきたプロフィール（以下、 $q_R$ と略称する）は、ユーザの要求を表すように適合文書の特徴を取り入れて更新されてきた。しかし、 $q_R$ に類似している文書を選択するだけでは、非適合文書も多く選択されてしまうということが明らかになっている。従って、プロフィールとの類似度が高いが実際には要求に適合していない非適合文書を選択しないようにすれば、精度を向上させることができると考えられる。そこで、過去に $q_R$ との類似度が高いと判断され選択された非適合文書の特徴づけるプロフィールを非適合プロフィール（以下、 $q_N$ と略称する）として作成する。 $q_N$ との類似度が高い文書は、過去に誤って選択された非適合文書集合に類似しているため、このような文書を選択しないようにすれば、従来手法では誤って選択されていた非適合文書の選択を回避することができる。

【0047】このような考えに基づく本発明の文書フィルタリング方法について更に具体的に説明する。

【0048】まず、従来のプロフィール更新同様、選択された文書から単語寄与度に基づき単語を抽出する。そして、抽出された単語が適合文書中の単語の場合には、単語に対するスコア $Score_{relN}(w_i)$ を式(16)により算出し、非適合文書中の単語の場合には $Score_{nrelN}(w_i)$ を式(17)によって算出する。

【0049】

【数9】

$$Score_{relN}(w_i) = wgt_{relN} \times Cont(w_i, q, d) \quad \dots (16)$$

$$Score_{nrelN}(w_i) = wgt_{nrelN} \times Cont(w_i, q, d) \quad \dots (17)$$

次に、上記の式によって求めた各単語のスコアを単語出現頻度 $t_f$ として扱い、以下の式(18)、(19)により各単語 $w_i$ の $TF \cdot IDF$ 値を算出する。

【0050】

【数10】



$$\text{Value}_{\text{relN}}(w_i) = \log(1 + \text{Score}_{\text{relN}}(w_i)) \times \log\left(\frac{M}{df_i}\right)$$

... (18)

$$\text{Value}_{\text{nrelN}}(w_i) = \log(1 + \text{Score}_{\text{nrelN}}(w_i)) \times \log\left(\frac{M}{df_i}\right)$$

... (19)

ただし、 $df_i$ は単語 $w_i$ が出現する文書数、 $M$ は $df$ のリスト作成に使用された文書数とする。

【0051】また、 $q_N$ および更新後の $q_N$ を式(2

$$\vec{q}_N = (q_{N1}, \dots, q_{Nn}) \quad \dots (20)$$

$$\vec{q}_{N\text{new}} = (q_{N1}', \dots, q_{Nn}') \quad \dots (21)$$

そして、 $q_R$ の更新とは逆に、抽出された単語 $w_i$ が適合文書中の単語の場合は、式(22)のように、 $q_N$ を表すベクトルから $\text{Value}_{\text{relN}}(w_i)$ を引き、非適合文書中の単語の場合は、式(23)のように、 $\text{Value}_{\text{nrelN}}(w_i)$ を加えることで、非適合文書の特徴を表すように $q_N$ を更新する。

$$q_{Ni}' = q_{Ni} - \text{Value}_{\text{relN}}(w_i) \quad \dots (22)$$

$$q_{Ni}' = q_{Ni} + \text{Value}_{\text{nrelN}}(w_i) \quad \dots (23)$$

そして、 $q_R$ との類似度が閾値を超えた文書について、 $q_N$ との類似度を計算し、 $q_N$ に対する閾値を超えた文書は過去に誤って選択した非適合文書に類似していると判断し、選択しない。

【0054】次に、図2に示すフローチャートを参照して、図1に示す実施形態における非適合プロファイルを利用した文書フィルタリングの処理手順について説明する。

【0055】まず、ユーザの要求である適合プロファイル $q_R$ と検索対象文書 $d$ との類似度 $\text{Sim}_R$ を計算し(ステップS11)、この計算した類似度 $\text{Sim}_R$ が所定の閾値 $\text{Thres}_R$ よりも大きいか否かを判定する(ステップS13)。計算した類似度 $\text{Sim}_R$ が所定の適合用閾値 $\text{Thres}_R$ より小さい場合には、最初のステップS11に戻り、次の検索対象文書に対して同じ処理を繰り返す。類似度 $\text{Sim}_R$ が所定の適合用閾値 $\text{Thres}_R$ よりも大きい場合には、非適合プロファイル $q_N$ と検索対象文書 $d$ との類似度 $\text{Sim}_N$ を計算する(ステップS15)。

【0056】この計算した類似度 $\text{Sim}_N$ が所定の非適合用閾値 $\text{Thres}_N$ より小さいか否かを判定する(ステップS17)。類似度 $\text{Sim}_N$ が所定の非適合用閾値

0)、式(21)で表すとする。

【0052】

【数11】

$\text{nrelN}(w_i)$ を加えることで、非適合文書の特徴を表すように $q_N$ を更新する。

【0053】

【数12】

$\text{Thres}_N$ より小さくない場合には、すなわち類似度 $\text{Sim}_N$ が所定の非適合用閾値 $\text{Thres}_N$ より大きい場合には、検索対象文書を非適合文書と見なして選択せず、すなわち除外し、最初のステップS11に戻り、次の検索対象文書に対して同じ処理を繰り返す。類似度 $\text{Sim}_N$ が所定の非適合用閾値 $\text{Thres}_N$ より小さい場合には、適合文書と見なして、ユーザに出力し、これに対する適合フィードバックを得て(ステップS19)、適合プロファイル $q_R$ および非適合プロファイル $q_N$ を上述した式(22)、(23)に示したように更新する(ステップS21)。以上の処理を全ての検索対象文書について繰り返す(ステップS23)。

【0057】以上の処理により適合プロファイル $q_R$ に類似していると判断された文書から、過去に誤って選択された非適合文書に類似している文書を削除し、選択された非適合文書の数を減少させることができる。

【0058】次に、図3に示すフローチャートを参照して、上述した図2の処理手順において非適合文書であると見なした文書をシステムにフィードバックするpseudo feedbackによる適合プロファイルを利用した文書フィルタリング処理について説明する。

【0059】図3に示す文書フィルタリング処理は、図

2に示した文書フィルタリング処理におけるステップS17での類似度 $Sim_N$ が所定の非適合用閾値 $Thres_N$ より小さいか否かについての判定において、類似度 $Sim_N$ が所定の非適合用閾値 $Thres_N$ よりも小さくない場合に、すなわち類似度 $Sim_N$ が所定の非適合用閾値 $Thres_N$ よりも大きい場合に、検索対象文書を非適合文書と見なして、pseudo feedbackし（ステップS31）、このpseudo feedbackによる情報で非適合プロファイル $q_N$ の更新を行うものである（ステップS33）。その他の作用は図2の作用と同じであり、図3において図2と同じ処理には同じステップ番号が付されている。

【0060】このように類似度 $Sim_N$ が所定の非適合用閾値 $Thres_N$ よりも大きいものを非適合文書と見なして、pseudo feedbackし、このpseudo feedbackによる情報も非適合プロファイルの更新に利用することで、非適合文書の更新に利用する情報を増大することができ、これにより非適合プロファイル $q_N$ との類似度の閾値を厳しくした際の選択文書の減少に伴う非適合プロファイル $q_N$ の更新に利用するフィードバック情報の減少を補うことができる。

【0061】すなわち、上述したように、pseudo feedbackを取り入れることにより、非適合プロファイル $q_N$ との閾値を厳しくすることができ、非適合プロファイル $q_N$ とのフィルタリングの効果を生かし、かつ多くの文書情報を基に有効な非適合プロファイル $q_N$ を作成することができる。

【0062】次に、図4に示すフローチャートを参照して、上述した図2および図3における非適合プロファイルの更新処理について説明する。

【0063】図4に示すように、ベクトル空間モデルによって表されたプロファイル $q$ および更新後のプロファイル $q_{new}$ 、および選択文書 $d$ に対して、まず選択文書 $d$ から単語集合 $W$ を抽出し（ステップS53）、選択文書 $d$ が適合文書であるか否かを判定する（ステップS57）。

【0064】そして、選択文書 $d$ が適合文書である場合には、各単語に対するスコアを式(16)のように算出し（ステップS58）、それから式(22)で示したと同様に各単語 $w_i$ のスコアをプロファイル $q$ から引く（ステップS61）。また、選択文書 $d$ が適合文書でない場合には、各単語に対するスコアを式(17)のように算出し（ステップS57）、式(23)で示したと同様に各単語 $w_i$ のスコアをプロファイル $q$ に加え（ステップS63）、このような加減算したプロファイル $q$ を更新後のプロファイル $q_{new}$ として設定する（ステップS65）。

【0065】

【発明の効果】以上説明したように、本発明によれば、ユーザの要求に適合しないと判定された非適合文書を特徴づけるプロファイルを非適合プロファイルとして作成し、適合プロファイルとの類似度が大きい場合、非適合プロファイルとの類似度を算出し、この類似度が大きい場合、検索対象文書を非適合文書と見なして除外するので、適合文書とともに選択される非適合文書の数を少なくすることができ、フィルタリング性能を向上することができる。

【0066】また、本発明によれば、選択文書の特徴を表す単語を抽出し、この抽出された単語の寄与度を算出し、単語寄与度に重みをかけてスコアを算出し、選択文書が適合文書である場合には、スコアを非適合プロファイルから減算し、選択文書が非適合文書である場合には、スコアを非適合プロファイルに加算して更新するので、非適合プロファイルを適確に更新して文書フィルタリング精度を向上することができる。

【図面の簡単な説明】

【図1】本発明の一実施形態に係る文書フィルタリング方法を実施する文書フィルタリング装置の構成を示すブロック図である。

【図2】図1に示す実施形態における非適合プロファイルを利用した文書フィルタリングの処理手順を示すフローチャートである。

【図3】図2に示した文書フィルタリング処理において非適合文書であると見なした文書をシステムにフィードバックするpseudo feedbackによる適合プロファイルを利用した文書フィルタリング処理を示すフローチャートである。

【図4】図1の実施形態における非適合プロファイルの更新処理を示すフローチャートである。

【図5】プロファイル更新を利用した従来の文書フィルタリング処理を示すフローチャートである。

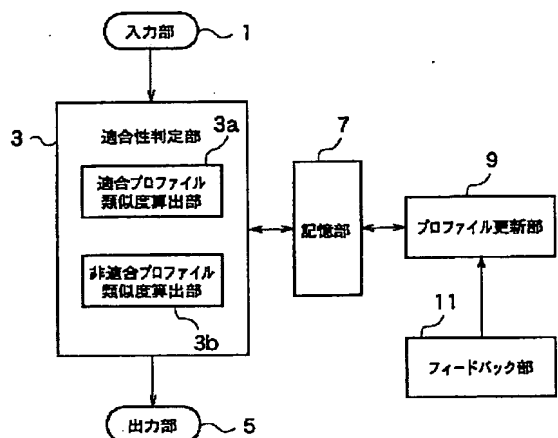
【図6】図5に示した文書フィルタリング処理におけるプロファイル更新処理を示すフローチャートである。

【図7】図5に示した従来の文書フィルタリング処理を評価する実験結果である適合文書と非適合文書の類似度を示すグラフである。

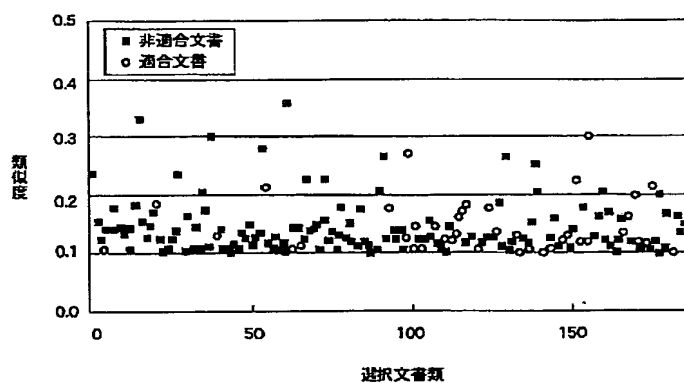
【符号の説明】

- 1 入力部
- 3 適合性判定部
- 3a 適合プロファイル類似度算出部
- 3b 非適合プロファイル類似度算出部
- 5 出力部
- 7 記憶部
- 9 プロファイル更新部
- 11 フィードバック部

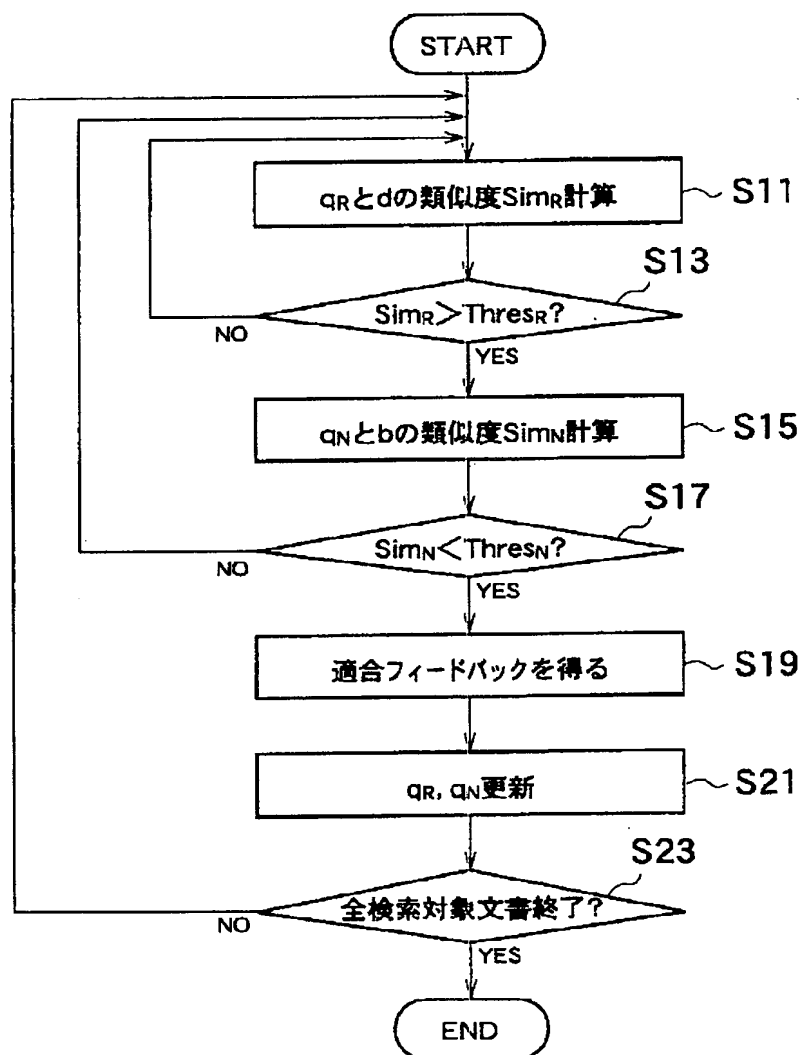
【図1】



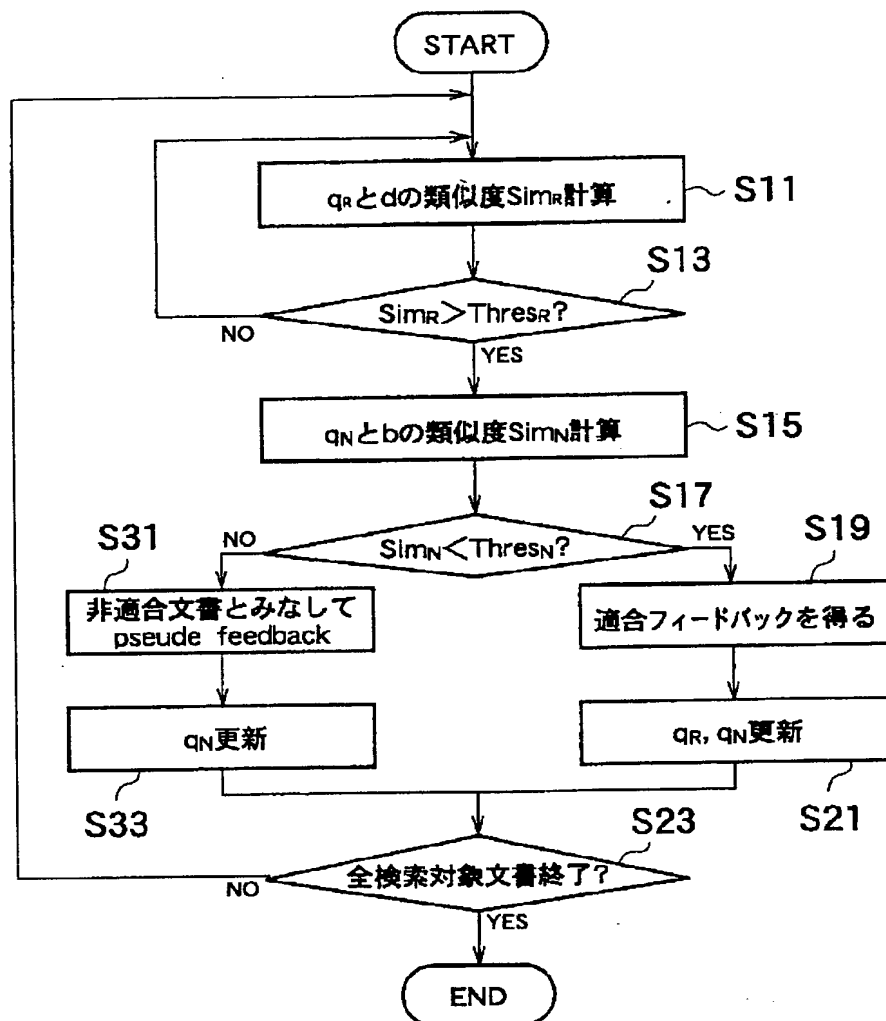
【図7】



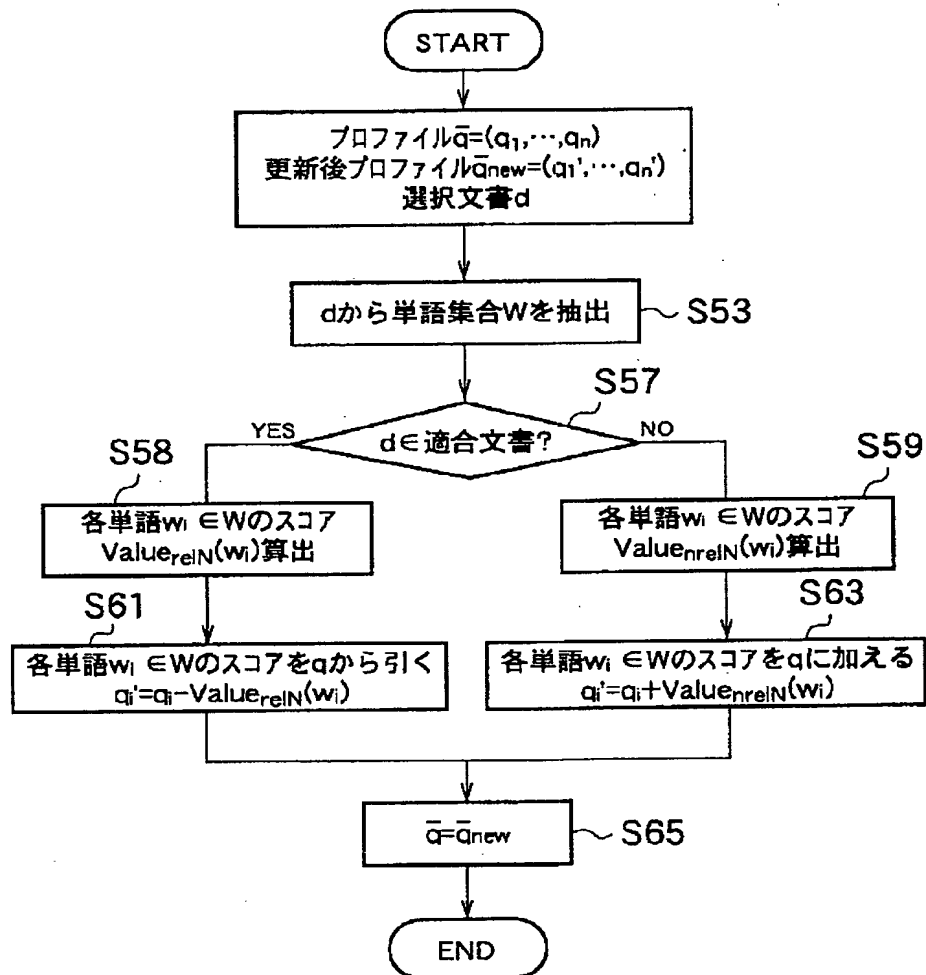
【図2】



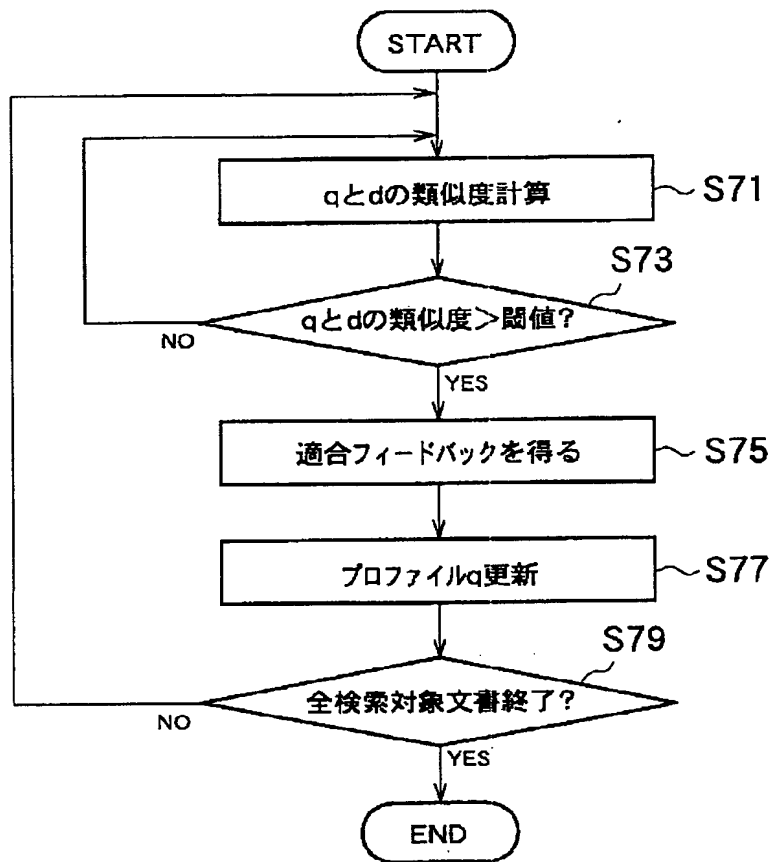
【図3】



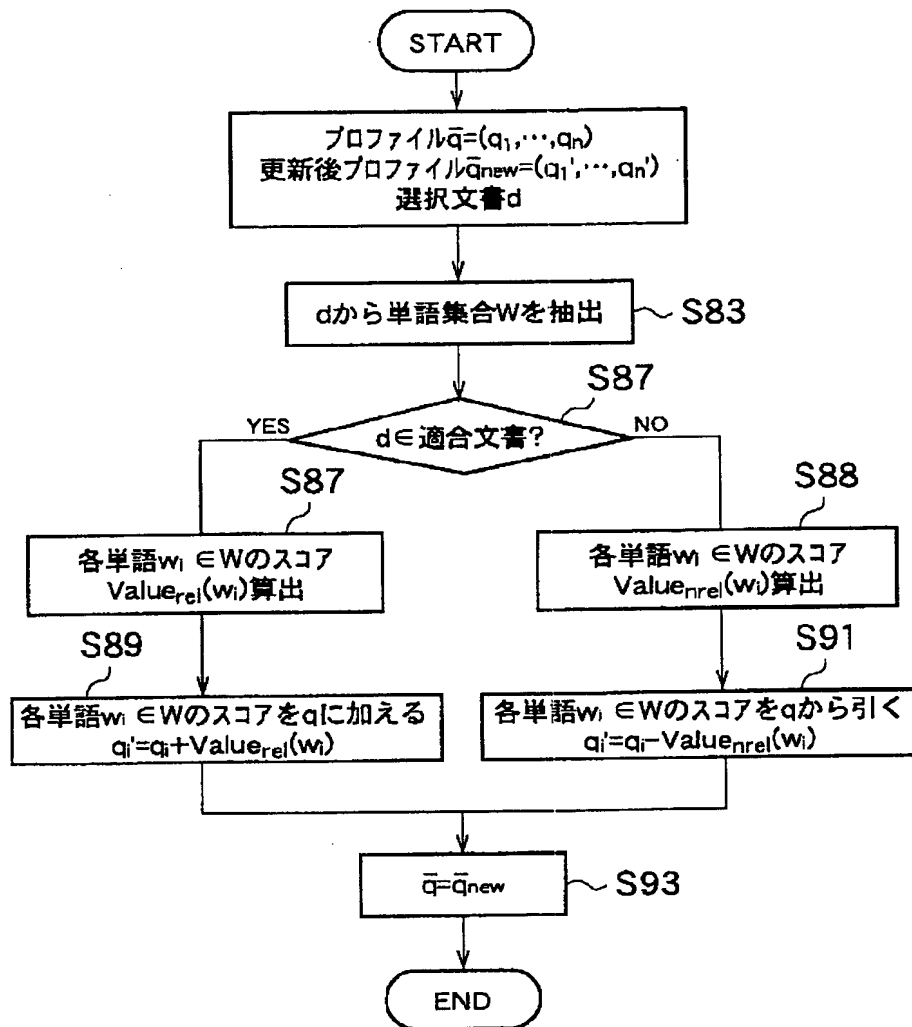
【図4】



【図5】



【図6】



フロントページの続き

(72)発明者 松本 一則  
埼玉県上福岡市大原2-1-15 株式会社  
ケイディディ研究所内

(72)発明者 橋本 和夫  
埼玉県上福岡市大原2-1-15 株式会社  
ケイディディ研究所内  
Fターム(参考) 5B075 ND03 PR06 PR08 QM08